Capstone Project Report

9th December, 2024
Team: CAP 103
Members: Mihir Upadhyay, Harshit Bhargava, Rishabh Patil

Introduction

This report explores patterns in professor ratings using data from RateMyProfessor.com, addressing questions related to gender bias, subject differences, and predictors of student evaluations. Through rigorous statistical tests and machine learning models, the analysis investigates factors influencing average ratings, difficulty, and the likelihood of receiving a "pepper." A significance threshold of α=0.005 was applied throughout to ensure the reliability of the findings.

Data Cleaning and Preprocessing

The data cleaning and preprocessing steps involved several key actions to ensure the dataset's quality and reliability. First, the dataset was filtered to include only rows where the male and female columns differed, eliminating redundant gender data. Professors with fewer than three ratings (the median number of ratings) were excluded to ensure sufficient data for meaningful analysis. Row-wise elimination was then applied to remove incomplete records. The dataset was subsequently split into two groups: ratings for male professors and ratings for female professors, specifically for significance testing. For certain analyses, additional transformations such as combining numerical ratings with tag data were performed. Outliers and skewness in the data were addressed using Box-Cox transformations.

Authors' Contributions (Collaborative Effort with Shared (not divided) Responsibilities)(Mihir Upadhyay, Harshit Bhargava, Rishabh Patil)

We collaboratively addressed each question through in-depth discussions to determine the most appropriate statistical tests and data preprocessing steps based on the dataset. For data cleaning, we ensured accuracy by filtering rows with null values, applying thresholds like requiring at least three ratings per professor, and handling missing or redundant columns such as female and class_again_prop. Visualizations, including violin plots, box plots, and KDEs, were created to better understand patterns and communicate findings effectively.

For modeling questions, we explored various approaches, such as linear regression, polynomial regression, and logistic regression, tailoring the methods to each task. For example, we addressed collinearity by dropping highly correlated variables like "Inspirational" and "Caring" for tag-based models and standardized predictors to ensure comparable coefficients. These efforts improved model performance, achieving high R^2 values for predicting average ratings.

Additionally, for the extra credit question, we examined whether a professor's subject influences ratings using the Kruskal-Wallis test, identifying significant differences across subjects. Subjects like "Languages," "Music," and "Communication" received the highest mean ranks, while "Economics," "Chemistry," and "Engineering" ranked the lowest.

Seeding

The random number generator is initially seeded using the N-Number of a team member, Rishabh Patil (16150234), to ensure the reproducibility and uniqueness of the results. Initially, a random integer between 0 and 100 was generated after seeding and displayed for verification. However, we ultimately decided to directly use the N-Number as the random state value instead of seeding.

1. Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to a low sample size –as small as n = 1 (Mitchell & Martin, 2018), failure to control for confounders such as teaching experience (Centra & Gaubatz, 2000) or obvious p-hacking (MacNell et al., 2015). We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset. Hint: A significance test is probably required

Hypotheses:

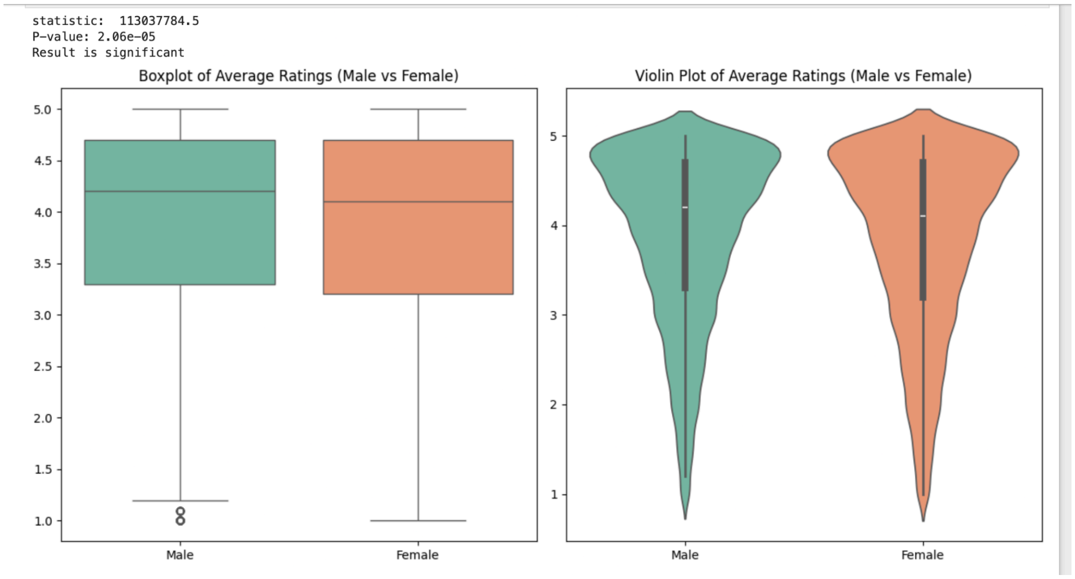
- **Null Hypothesis:** There is no pro-male gender bias in student ratings. Any observed difference between the average ratings for male and female professors is not statistically significant.
- **Alternative Hypothesis:** There is a pro-male gender bias in student ratings. The average ratings for male professors are significantly higher than those for female professors.

D: The dataset was filtered to include only rows where the male and female columns differ, with professors having at least three ratings (the median number of ratings). Row-wise elimination was applied, and the data was then split into two groups: ratings for male professors and ratings for female professors. A one-sided Mann-Whitney U test was used to compare the ratings between these two groups.

Y: The dataset was filtered to include only rows where the male and female columns differ, eliminating redundant gender data. Rows with fewer than three ratings were excluded to ensure data quality, as the average rating is more meaningful with a higher number of ratings. The Mann-Whitney U test was chosen for its robustness as a non-parametric method, ideal for comparing two independent groups with ordinal and nominal data. A one-sided Mann-Whitney U test was performed, testing the alternative hypothesis that male professors have higher ratings (alternative='greater').

F: The U-statistic is **113037784.5**, accompanied by a p-value of **2.06e-05**, indicating a significant difference between the two groups.

A: Since the P-value is below the alpha level of 0.005, the null hypothesis is rejected. With the low P-value, it can be concluded that male professors receive higher average ratings than female professors (pro-male bias), indicating the presence of a pro-male gender bias in the dataset.



```
statistic:  113037784.5
P-value: 2.06e-05
Result is significant
```

Boxplot of Average Ratings (Male vs Female)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Male
Female
Violin Plot of Average Ratings (Male vs Female)
5
4
3
2
1
Male
Female

2. Is there a gender difference in the spread (variance/dispersion) of the ratings distribution? Again, it is advisable to consider the statistical significance of any observed gender differences in this spread.

Hypotheses:

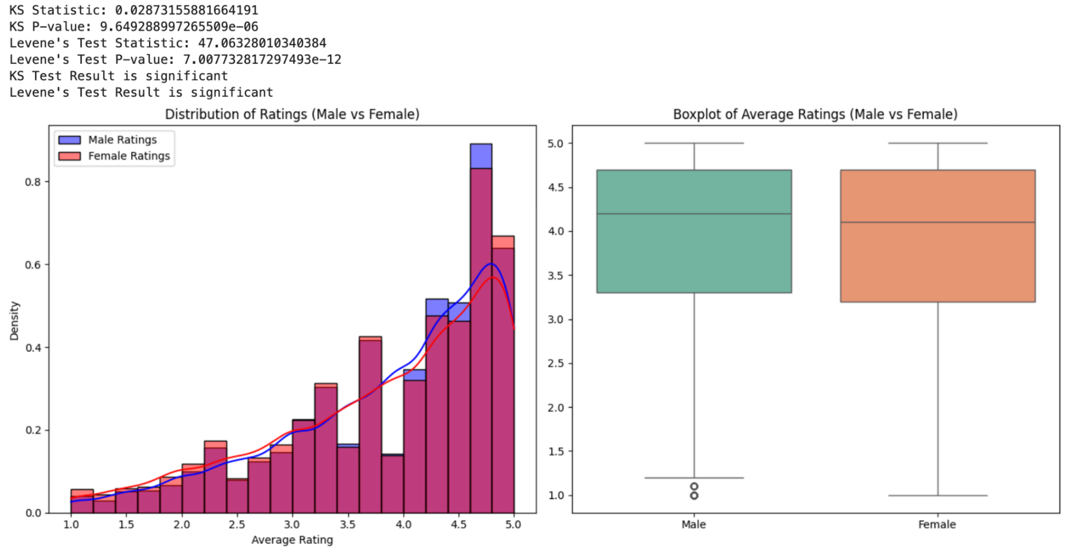
- **Null Hypothesis:** There is no statistically significant difference between the variances (spread) of the ratings distributions for male and female professors. Any observed differences are due to random chance.
- **Alternative Hypothesis:** There is a statistically significant difference between the variances of the ratings distributions for male and female professors. The observed differences are unlikely to be due to random chance.

D: The dataset was filtered to include only rows where the male and female columns differ, with professors having at least three ratings (the median number of ratings). Row-wise elimination was then applied. The data was split into two groups: ratings for male professors and ratings for female professors. The Kolmogorov-Smirnov (KS) Test and Levene's Test were performed on the ratings of these two groups.

Y: The dataset was filtered to include only rows where the male and female columns differ, eliminating redundant gender data. Rows with fewer than three ratings were excluded to ensure data quality and reliability. The KS Test was performed, as it is a non-parametric method that compares the cumulative distributions of two groups, making it suitable for detecting differences in overall shape or spread. While the KS Test showed significance, it does not specify whether the difference is due to spread or other factors, which is why Levene's Test was also conducted to specifically assess differences in variance, providing a direct comparison of the data spread.

F: The KS-statistic is **0.0287**, accompanied by a p-value of **9.65e-06**. Levene's test statistic is **47.063**, accompanied by a p-value of **7.01e-12**, both indicating a significant difference between the two groups.

A: Kolmogorov-Smirnov Test: A significant difference was found, indicating that the ratings distributions for male and female professors are different. **Levene's Test**: Result: A significant difference was also observed, showing that the variances in the ratings distributions for male and female professors differ. Both tests revealed statistically significant differences (P-value < 0.005) in both the distributions and variances of ratings for male and female professors, suggesting that not only do male and female professors have different rating patterns, but they also exhibit differing spreads in their rating distributions.



```
KS Statistic: 0.02873155881664191
KS P-value: 9.649288997265509e-06
Levene's Test Statistic: 47.06328010340384
Levene's Test P-value: 7.007732817297493e-12
KS Test Result is significant
Levene's Test Result is significant
```

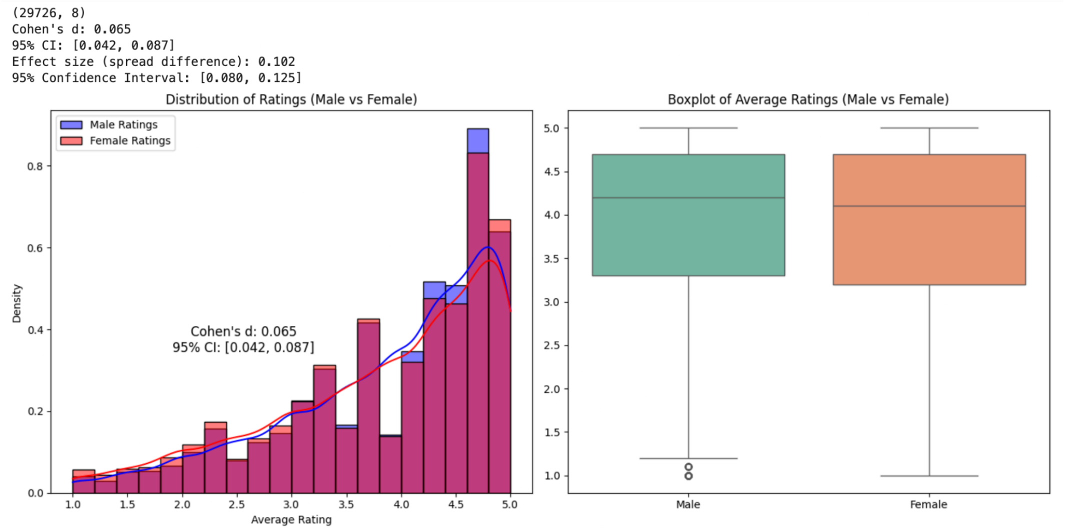
3. What is the likely size of both of these effects (gender bias in average rating, gender bias in spread of average rating), as estimated from this dataset? Please use 95% confidence and make sure to report each/both.

D: The dataset was filtered to include only rows where the male and female columns differ, with professors having at least three ratings (the median number of ratings). Row-wise elimination was applied, and the data was then split into two groups: ratings for male professors and ratings for female professors. Cohen's d was calculated along with a 95% confidence interval (CI) for Cohen's d, as well as Cohen's d and 95% CI for the spread difference effect size.

Y: The dataset was filtered to include only rows where the male and female columns differ, eliminating redundancy in gender data. Rows with fewer than three ratings were excluded to maintain data quality and reliability. Cohen's d was used to quantify the magnitude of the difference in average ratings, providing a measure of effect size. The spread difference was assessed using variance-based effect size, as variance directly reflects the dispersion in ratings. Confidence intervals were calculated to offer a range of likely values for both effect sizes.

F: Gender Bias in Average Rating: Cohen's d: 0.065 and **95% CI:** [0.042, 0.087], **Gender Bias in Spread of Average Rating**: **Effect Size (Spread Difference):** 0.102, **95% CI:** [0.080, 0.125]

A: The gender bias in **average ratings** shows a small effect size (d=0.065), suggesting minimal differences between male and female professors' ratings. The gender bias in the **spread of ratings** also shows a small effect size (d=0.102), indicating slightly more variation in ratings for one group.



```
(29726, 8)
Cohen's d: 0.065
95% CI: [0.042, 0.087]
Effect size (spread difference): 0.102
95% Confidence Interval: [0.080, 0.125]
```

4. Is there a gender difference in the tags awarded by students? Make sure to teach each of the 20 tags for a potential gender difference and report which of them exhibit a statistically significant difference. Comment on the 3 most gendered (lowest p-value) and least gendered (highest p-value) tags.

Hypotheses:
For each tag, the following hypotheses were tested:

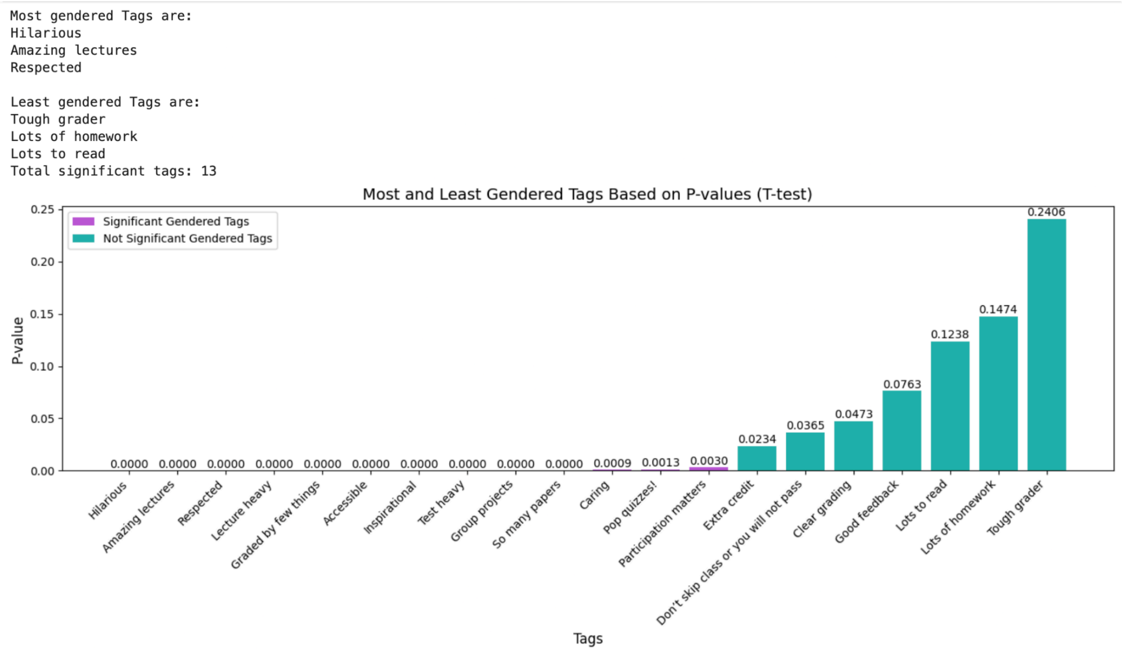
- **Null Hypothesis:** There is no gender difference in the frequency of the tag awarded to male and female professors.
- **Alternative Hypothesis:** There is a gender difference in the frequency of the tag awarded to male and female professors.

D: Combined numerical ratings data and tag data into a single dataset and Filtered the dataset to include only rows where the male and female columns differ. Row-wise elimination was then performed on the dataset. The data is then split into two groups: ratings for male professors and ratings for female professors. Performed a Welch's t-test for each of the 20 tags and then Counted the number of tags with statistically significant gender differences.

Y: The dataset was filtered to include only rows where the male and female columns differ, ensuring that the analysis focused on gender comparisons. Rows with fewer than three ratings were excluded to ensure data quality, as tags awarded to professors with very few ratings may not be representative. This approach improves the reliability of the results by including only professors with sufficient data for meaningful analysis. Welch's t-test was chosen because it is appropriate for comparing the means of two independent groups when their variances are unequal, which is often the case with real-world data. This test also does not assume equal sample sizes, making it more robust when comparing groups with differing numbers of ratings. By applying Welch's t-test, the analysis accounts for possible differences in variance, providing a more accurate comparison of gender differences in tag frequency.

F: Total significant tags: 13 out of 20 tags, **Most Gendered Tags (smallest p-values):** Hilarious, Amazing lectures, Respected**, Least Gendered Tags (largest p-values):** Tough grader, Lots of homework, Lots to read

A: There are significant gender differences in 13 out of the 20 tags awarded by students, with tags like "Hilarious," "Amazing lectures," and "Respected" being the most gendered tags. Conversely, tags such as "Tough grader," "Lots of homework," and "Lots to read" were the least gendered tags.



```
Most gendered Tags are:
Hilarious
Amazing lectures
Respected

Least gendered Tags are:
Tough grader
Lots of homework
Lots to read
Total significant tags: 13
```

5. Is there a gender difference in terms of average difficulty? Again, a significance test is indicated.

Hypotheses:

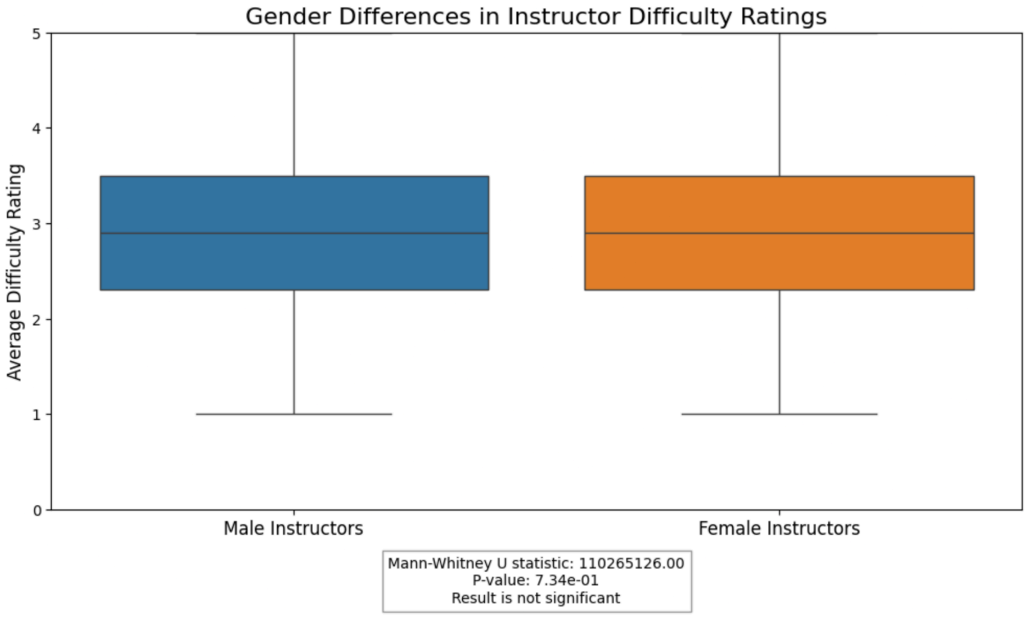
- **Null Hypothesis:** There is no significant difference in the average difficulty ratings between male and female professors.
- **Alternative Hypothesis:** There is a significant difference in the average difficulty ratings between male and female professors.

D: The dataset was filtered to include only rows where the male and female columns differ, ensuring that only professors with at least three ratings (the median number of ratings) were included. Rows with insufficient data were removed. The dataset was then divided into two groups: ratings for male professors and ratings for female professors. These two groups were compared using a two-sided Mann-Whitney U test.

Y: The dataset was filtered to include only rows where the male and female columns differ, eliminating redundant gender data. Rows with fewer than three ratings were excluded to ensure data quality, as the average rating is more meaningful with a higher number of ratings. The Mann-Whitney U test was chosen for its robustness as a non-parametric method, ideal for comparing two independent groups with ordinal and nominal data. A two-sided Mann-Whitney U test was performed, testing the alternative hypothesis of whether male professors have different difficulty ratings than female professors (alternative='two-sided'').

F: The U-statistic is **110265126.0**, accompanied by a p-value of **0.734**, indicating an insignificant difference between the two groups.

A: Result: No Significant difference - There is no difference in the average difficulty ratings between male and female professors. Since the P-value is greater than the significance (0.005), we fail to reject the null hypothesis that the differences seen in ratings are due to chance and conclude that there is no difference in the average difficulty ratings between male and female professors.



Mann-Whitney U statistic: 110265126.00
P-value: 7.34e-01
Result is not significant

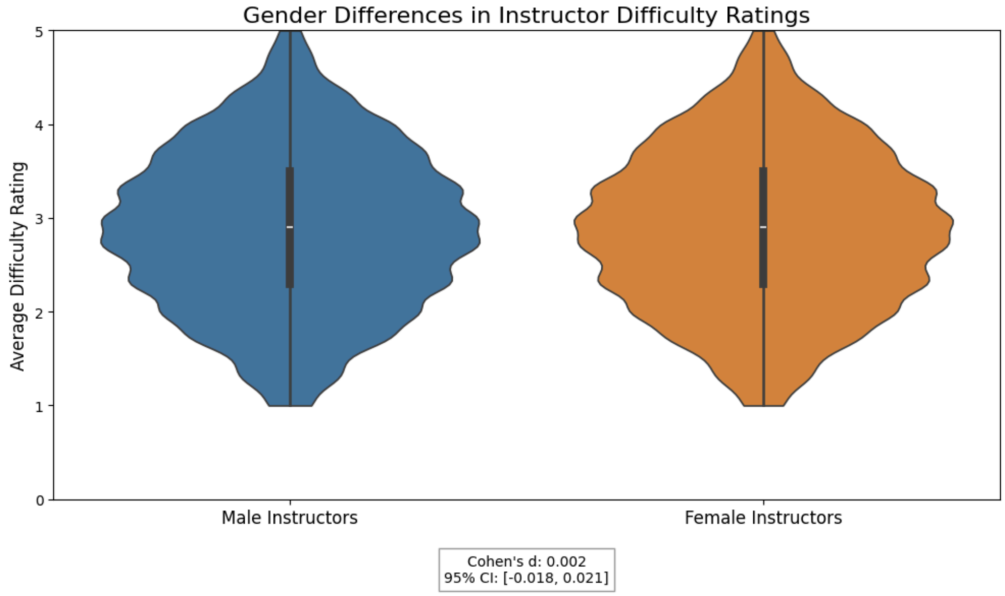
6. Please quantify the likely size of this effect at 95% confidence

D: The dataset was filtered to include only rows where the male and female columns differ, ensuring that professors with at least three ratings (the median number of ratings) were included. Rows with insufficient data were removed. The data was then divided into two groups: ratings for male professors and ratings for female professors. Cohen's d was calculated, along with a 95% confidence interval (CI) for the effect size.

Y: Filtered the dataset to include only rows where the male and female columns differ, this ensures that is no redundancy in gender data. Rows with fewer than three ratings were excluded to ensure data quality and reliability. Cohen's d is used to quantify the magnitude of the difference in average ratings in order to measure the effect size. Confidence intervals were calculated to provide a range of likely values for effect size.

F: Cohen's d: 0.002 and **95% CI:** [-0.018, 0.021]

A: The likely size of the gender difference in average difficulty ratings is extremely small (Cohen's d=0.002), and the 95% confidence interval ([-0.018, 0.021]) includes zero, indicating that the true effect is likely insignificant. This indicates that there is no meaningful difference in perceived difficulty based on the professor's gender.



Cohen's d: 0.002
95% CI: [-0.018, 0.021]

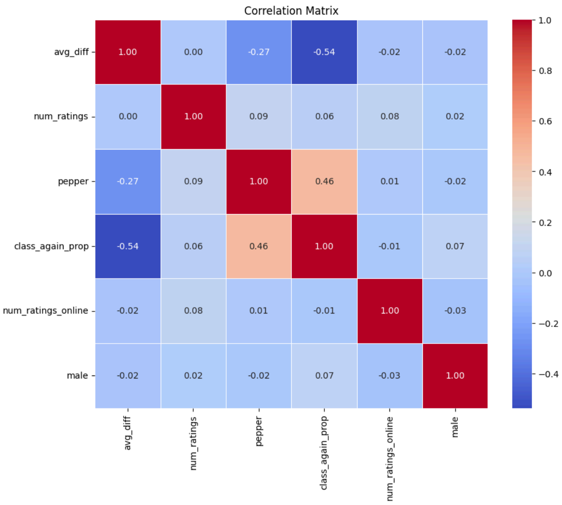
7. Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R² and RMSE of this model. Which of these factors is most strongly predictive of average rating? Hint: Make sure to address collinearity concern

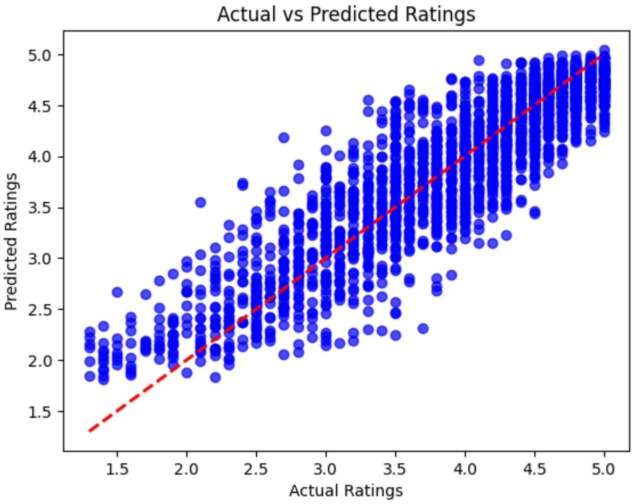
D: Preprocessed the Dataset by dropping the female column and row-wise elimination was performed on the dataset. The numerical predictors (excluding male and pepper) were standardized and collinearity was examined using a correlation heatmap. Built and trained a linear regression model using the scaled numerical predictors to predict average ratings. Coefficients, R², and RMSE were computed, and the most predictive variables were identified based on the regression coefficients.

Y: The female column was dropped to reduce redundancy, as including both gender columns would introduce multicollinearity, making coefficients unreliable. Standard scaling was applied to handle different predictor ranges, ensuring the coefficients were comparable. Linear regression was chosen because it is well-suited to continuous target variables and interpretable in terms of feature importance.

F: *Linear Regression:* R² = 0.805 and RMSE= 0.365, The most predictive factors were **"The proportion of students that said they would take the class again", (Coefficient = 0.627)**, **"Received a "pepper" (Coefficient = 0.203),** and **male (Coefficient = 0.04).**

A: Given that the R² value is high (explaining 81.8% of the variance in average ratings) and the RMSE is low, the regression model predicts average ratings accurately. The most strongly predictive factor is **"The proportion of students that said they would take the class again"**, suggesting that the likelihood of a student taking the same class again significantly influences average ratings. A limitation is that the model assumes linearity, which might not capture complex relationships between predictors and the target variable.





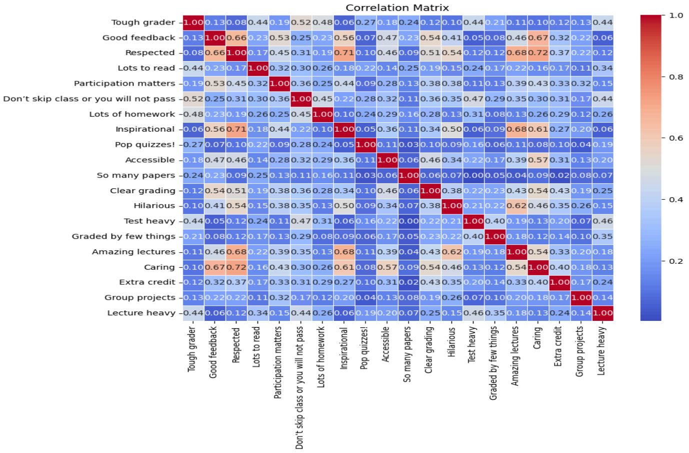
8. Build a regression model predicting average ratings from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R² and RMSE of this model. Which of these tags is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns. Also comment on how this model compares to the previous one.

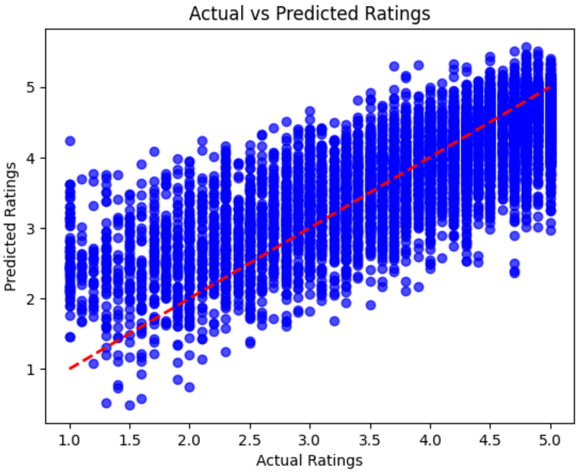
D: Preprocessed the dataset by Combining numerical and tag-based features into a single dataset and retained only relevant columns(tag-based features) for tag-based predictors and the target variable (avg_rating). Filtered the dataset for professors with at least three ratings (median number of ratings). Row-wise elimination was then performed on the dataset. The collinearity was examined using a correlation heatmap and the highly correlated columns ("Inspirational" and "Caring") with correlation thresholds above 0.7 were dropped. Applied Box-Cox transformations on the columns and Created polynomial features (up to degree 2). Built and trained a Polynomial regression model using the transformed data to predict average ratings. Coefficients, R², and RMSE were computed, and the most predictive variables were identified based on the regression coefficients (absolute value).

Y: Filtering for professors with at least three ratings enhanced the reliability of the data. Multicollinearity was addressed to ensure the coefficients accurately reflected true predictive relationships. Box-Cox transformation was applied to improve the skewness of predictors, and polynomial features were created to capture potential interactions between tags. Polynomial regression was chosen because the relationships between features might not be linear, and for this model, the results from polynomial regression were better than those from linear regression. However, the results of this model were not better than those from the previous question.

F: ***Linear Regression:*** **R^2 = 0.628 and RMSE= 0.603.** The most predictive factors were **“Tough Grader”: Coefficient = -0.632**, **“Good Feedback”: Coefficient = 0.465,** and **“Respected”: Coefficient = 0.429**

A: The model using only tags explained 62.7% of the variance in average ratings with a moderate RMSE. The most predictive tag was **“Tough Grader”**, which had a significant negative impact on ratings, followed by positive contributions from **“Good Feedback”** and **“Respected”.** Compared to the previous model (Q7), which achieved R² = 0.818 and RMSE = 0.368, the tags-only model performed worse, indicating that numerical factors have stronger predictive power than tags alone. Limitations include the exclusion of numerical predictors, which may limit the scope of this model's applicability. Integrating tags and numerical factors could lead to better performance of the model.

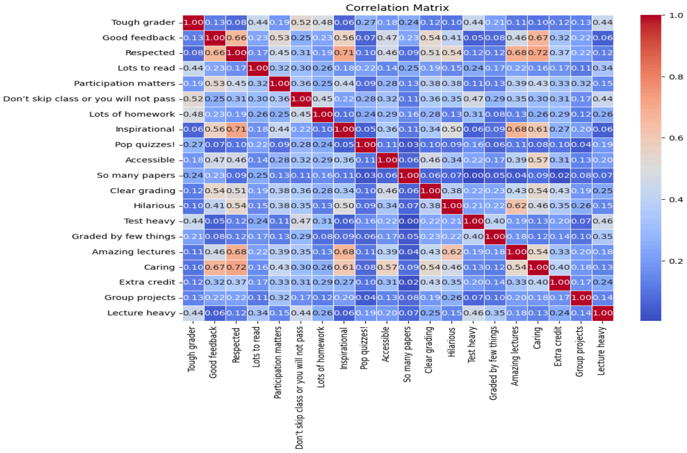


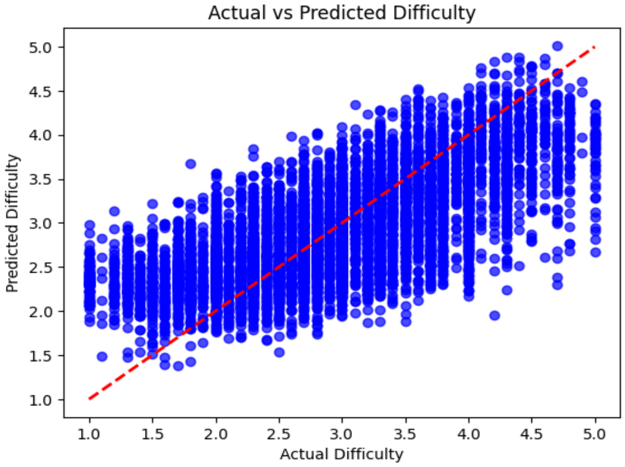


9. Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty? Hint: Make sure to address collinearity concern

D: Preprocessed the dataset by Combining numerical and tag-based features into a single dataset and retained only relevant columns(tag-based features) for tag-based predictors and the target variable (avg_difficulty). Filtered the dataset for professors with at least three ratings (median number of ratings). Row-wise elimination was then performed on the dataset. The collinearity was examined using a correlation heatmap and the highly correlated columns (“Inspirational” and “Caring”) with correlation thresholds above 0.7 were dropped. Applied Box-Cox transformations on the columns and Created polynomial features (up to degree 2). Built and trained a Polynomial Regression model using the transformed data to predict average ratings. Coefficients, R², and RMSE were computed, and the most predictive variables were identified based on the regression coefficients (absolute value).

Y: Filtering for professors with at least three ratings enhanced the reliability of the data. Multicollinearity was addressed to ensure the coefficients accurately reflected true predictive relationships. Box-Cox transformation was applied to improve the skewness of predictors, and polynomial features were created to capture potential interactions between tags. Polynomial regression was chosen because the relationships between features might not be linear, and for this model, the results from polynomial regression were better than those from linear regression.





F: ***Linear Regression:* R^2 = 0.49 and RMSE= 0.605,** The most predictive factors were **"Tough Grader": Coefficient = 0.817 (strong positive impact)**, **"Test heavy": Coefficient = 0.253 (moderate positive impact).,** and **"Clear Grading": Coefficient = -0.219 (moderate negative impact).**

A: The model explained 48.7% of the variance in average difficulty, with **"Tough Grader"** being the most strongly predictive tag, positively influencing difficulty perceptions. While the model captures some variation, it performed worse compared to the average rating prediction models (Q7 and Q8). A limitation is the reliance on tags alone, which might not fully capture other influential factors. Future improvements could include integrating numerical predictors to improve predictive power.

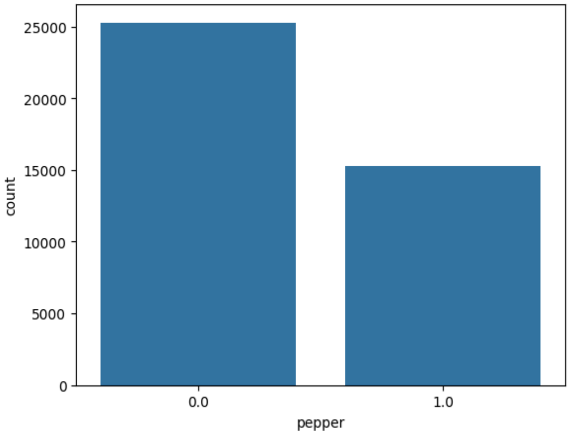
10. Build a classification model that predicts whether a professor receives a "pepper" from all available factors(both tags and numerical). Make sure to include model quality metrics such as AU(RO)C and also address class imbalance concerns.

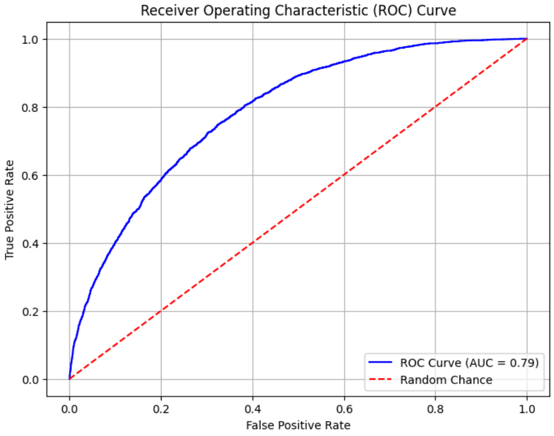
D: Preprocessed the dataset by Combining numerical and tag-based features into a single dataset. Dropped the "female" column and "class_again_prop" column and then filtered the dataset for professors with at least three ratings (median number of ratings. Row-wise elimination was then performed on the dataset. Scaled the predictors (excluding "male") using Standard scaling to standardize feature ranges. Built and trained a logistic regression model using scaled predictors to predict whether a professor received a "pepper." Evaluated model performance using accuracy, a confusion matrix, a classification report, and the AUC-ROC metric.

Y: The female column was dropped to reduce redundancy, as including both gender columns would introduce multicollinearity and the "class_again_prop column was dropped due to its high proportion of missing values (>80%). Keeping this column would have drastically reduced the dataset size during row-wise elimination. Filtering for professors with at least three ratings ensured reliable data, which reduced the number of rows but also helped counter class imbalance (Imbalance Ratio: ~1.66). Standard scaling was applied to handle different predictor ranges, ensuring the coefficients were comparable. Logistic regression was chosen for its simplicity, interpretability, and suitability for binary classification tasks.

F: ***Logistic Regression*: Accuracy = 0.71 (71%), AUC-ROC = 0.79**

A: The model performed moderately well in predicting whether a professor received a "pepper," with 72% accuracy AUC-ROC = 0.79 showing decent predictive power. However, the logistic regression assumes linear relationships, which may not capture all complexities in the data.





Extra Credit

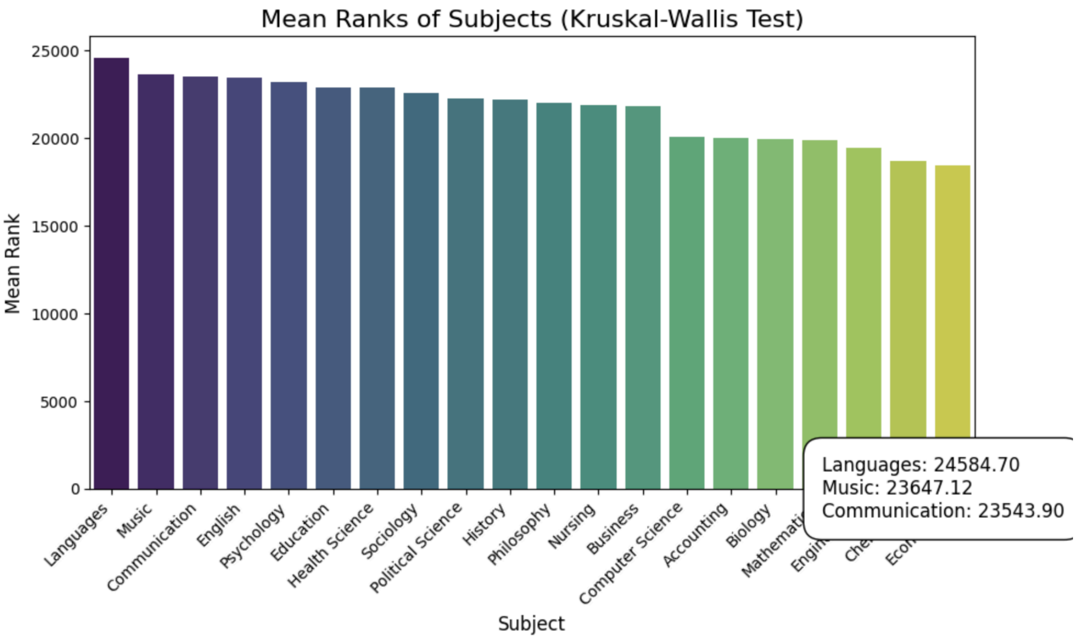
11. Is there a significant difference in average ratings across subjects?

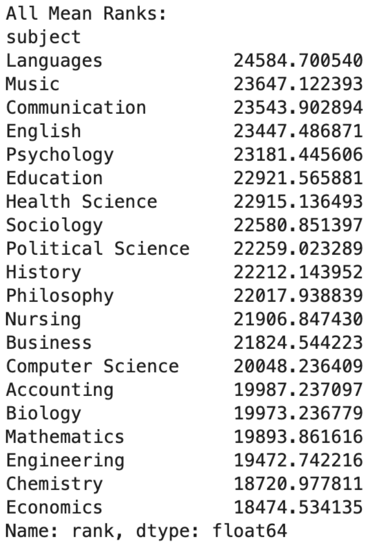
D: Combined subject data with average ratings into a single data frame. Dropped rows with missing avg_rating values and Filtered for subjects with more than 1,000 records. Performed Kruskal-Wallis test across all filtered subjects. Calculated mean ranks for each subject and sorted them to identify subjects with the highest and lowest mean ranks.

Y: Filtering Subjects with >1,000 Records ensures each subject has sufficient data to provide reliable comparisons. The Kruskal-Wallis test was chosen for its robustness in comparing multiple groups without assuming normality, which is suitable given the potential variability in rating distributions. The Mean ranks provide a clear, interpretable measure of how subjects differ relative to one another.

F: Kruskal-Wallis Test Results: **Statistic**: 832.80 and **P-value:** 2.29e-164 (highly significant). **Top 3 Subjects with Highest Mean Ranks:** Languages, Music, Communication, **Bottom 3 Subjects with Lowest Mean Ranks**: Economics, Chemistry, Engineering

A: There are significant differences in average ratings across subjects, with Languages, Music, and Communication having the highest mean ranks. These subjects are likely perceived more favourably by students. Conversely, Economics, Chemistry, and Engineering have the lowest mean ranks, indicating less favourable perceptions.





```
All Mean Ranks:
subject
Languages              24584.700540
Music                  23647.122393
Communication          23543.902894
English                23447.486871
Psychology             23181.445606
Education              22921.565881
Health Science         22915.136493
Sociology              22580.851397
Political Science      22259.023289
History                22212.143952
Philosophy             22017.938839
Nursing                21906.847430
Business               21824.544223
Computer Science       20048.236409
Accounting             19987.237097
Biology                19973.236779
Mathematics            19893.861616
Engineering            19472.742216
Chemistry              18720.977811
Economics              18474.534135
Name: rank, dtype: float64
```

```
Subjects: ['Computer Science', 'Education', 'Languages', 'Political Science', 'Business', 'Communication', 'Economics', 'Sociology', 'Engli
sh', 'Biology', 'Music', 'Mathematics', 'Psychology', 'Health Science', 'Chemistry', 'History', 'Philosophy', 'Engineering', 'Nursing', 'Ac
counting']
Statistic: 832.7950625183382
p-value: 2.2855521126063937e-164

Top 3 Subjects with the Highest Effect:
subject
Languages        24584.700540
Music            23647.122393
Communication    23543.902894
Name: rank, dtype: float64
```